Testing Measurement Invariance with Ordinal Missing Data: A Comparison of Estimators and Missing Data Techniques

Po-Yi Chen
Department of Psychology, University of Kansas;
Wei Wu
Department of Psychology, Indiana University-Purdue University Indianapolis
Mauricio Garnier-Villarreal
Marquette University
Benjamin Arthur Kite
Department of Psychology, University of Kansas
Fan Jia
University of Kansas

Abstract

Ordinal missing data are common in measurement equivalence/invariance (ME/I) testing studies. However, there is a lack of guidance on the appropriate method to deal with ordinal missing data in ME/I testing. Five methods may be used to deal with ordinal missing data in ME/I testing, including the continuous full information maximum likelihood estimation method (FIML), continuous robust FIML (rFIML), FIML with probit links (pFIML), FIML with logit links (lFIML), and mean and variance adjusted weight least squared estimation method combined with pairwise deletion (WLSMV_PD). The current study evaluates the relative performance of these methods in producing valid chi-square difference tests ($\Delta\chi 2$) and accurate parameter estimates. The result suggests that all methods except for WLSMV_PD can reasonably control the type I error rates of $\Delta\chi 2$ tests and maintain sufficient power to detect noninvariance in most conditions. Only pFIML and lFIML yield accurate factor loading estimates and standard errors across all the conditions. Recommendations are provided to researchers based on the results.

Keywords: Measurement invariance, missing data, ordinal data analysis

---

This is the author's manuscript of the article published in final edited form as:

Chen, P.-Y., Wu, W., Garnier-Villarreal, M., Kite, B. A., & Jia, F. (2020). Testing Measurement Invariance with Ordinal Missing Data: A Comparison of Estimators and Missing Data Techniques. Multivariate Behavioral Research, 55(1), 87–101.
https://doi.org/10.1080/00273171.2019.1608799

Introduction

Measurement equivalent/invariance (ME/I) is an important and desirable property for psychological tests or scales (Brown, 2006). ME/I concerns whether the relationships among observable indicators and underlying latent constructs are identical across groups (Millsap, 2011). ME/I is typically tested through a multistep process using multiple group confirmatory factor analysis (MG-CFA) in the structural equation modeling (SEM) framework. This process involves a series of chi-square difference (^{2}) tests between nested MG-CFA models, through which the level of ME/I can be established (described in detail below).

Given the predominant use of Likert-type scales in the social and behavioral sciences, the indicators are often ordinal in nature. In addition, missing data are also likely to occur. Previous studies have shown that problems with either ordinal or missing data can affect ME/I tests using MG-CFA (e.g., Sass, Schmitt, & Marsh, 2014; Widaman, Grimm, Early, Robins, & Conger, 2013). However, few studies have considered the case where the two problems co-exist.

At least five methods may be used for ME/I testing with the presence of ordinal missing data. We refer to them as continuous full information likelihood method (FIML), robust continuous full information likelihood method (rFIML), full information likelihood method with probit links (pFIML), full information likelihood method with logit links (lFIML), and the mean and variance adjusted weight least squared estimation method (WLSMV) combined with pairwise deletion (WLSMV_PD). These methods all have their strengths and limitations. Briefly speaking, the four FIML-based methods handle missing data directly, however, they either assume that the ordinal data are continuous (FIML and rFIML) or cannot include auxiliary variables (missing data predictors that are not part of the tested model) into the analyses due to computational burden (pFIML and lFIML).

WLSMV_PD, conversely, accounts for the ordinal nature of the data and auxiliary variables simultaneously. However, pairwise deletion is not an ideal method to deal with missing data. Note that multiple imputation, a more advanced missing data technique, may be combined with WLSMV to deal with missing ordinal data (Teman, 2012). Limited research in the past has shown that multiple imputation combined with WLSMV will produce accurate parameter and standard error estimates (Asparouhov & Muthén, 2010a; Teman, 2012). However, there is so far no good way to pool the ^{2} test statistics across the imputed data sets when WLSMV is used (Liu et al., 2017). Given that ^{2} tests are critical for ME/I testing, this is an indisputable limitation. Thus, we did not include this combination in the current study.

Considering that more than one method may be used to deal with ordinal missing data in ME/I testing and none of them seem to be ideal in theory, the question natually arises as to which method should be preferred. A guidance for this issue would be helpful for empirical researchers. Thus, the purpose of the study is to evaluate

the relative performances of these five methods on producing accuracy $\Delta\chi^2$ tests for ME/I testing, accurate parameter estimates, and standard errors using a simulation study.

The rest of the article is organized as follows. The typical process of ME/I testing using MG-CFA was reviewed first, followed by a description of the five methods to deal with missing ordinal data in ME/I testing. The design and results of the simulation study are then presented. An empirical example is also provided to illustrate the five methods. Based on the simulation results, practical recommendations to researchers are provided. The article is concluded by a discussion of limitations of the current study and directions for future research.

A typical ME/I testing process

There are four commonly tested invariance models: configural, metric, scalar, and strict invariance models, representing four levels of measurement invariance from least restricted to most restricted, respectively. In the configural invariance model, the factorial patterns (i.e., the patterns of free and fixed factor loadings) are assumed to be equal across groups. The metric invariance model is a configural model plus equivalent factor loadings across groups. The scalar invariance model goes beyond the metric model by further assuming equivalent thresholds across groups, and the strict invariance model is the scalar model plus equivalent residual variances across groups.

A typical ME/I testing process involves comparing the above four models in sequence to determine which level of measurement invariance is achieved (Kline, 2015). A configural invariance model is tested first in terms of its model fit. With a sufficient model fit of the configural invariance model, the metric invariance model is tested against the configural model using a $\Delta\chi^2$ test. If the $\Delta\chi^2$ test is not significant, indicating that adding equality constraints on factor loadings will not cause a significant decrease in model fit, then, the metric invariance model is retained. Researchers can further test the scalar invariance model against the metric invariance model using a $\Delta\chi^2$test, and so forth.

Note that when indicators are ordinal, past research suggested skipping the metric invariance model and directly comparing the configural model to the scalar invariance model. The rationale is that the probability of an individual endorsing a certain category in an item is jointly determined by the factor loadings and thresholds, which makes it reasonable to test the two sets of parameters simultaneously (Sass et al., 2014). We, thus, followed the suggestion in our study.

Methods to deal with ordinal missing data in ME/I tests

The past research on ME/I testing in SEM has been focused on complete and continuous data (e.g., Meredith, 1993). Only until recently, attention has been paid to issues raised by missing data (e.g., Widaman et al., 2013). As aforementioned, five methods may be used for ME/I testing with the presence of missing data. These methods are described in detail below.

Continuous full information maximum likelihood

Full information maximum likelihood method is one of most effective methods for handling missing data in SEM for ignorable missingness (Enders, 2010). It accounts for missing data by allowing case-wise log likelihood functions tailored according to the missing data pattern for each case and can use all the available information in the data simultaneously. Although, likelihood functions can be written based on different distributional assumptions; the typical one is built based on the multivariate normal assumption in SEM. We refer to this method as FIML in the current study. Despite that FIML assumes multivariate normality, it is still used in practice for ordinal indicators if it is reasonable to treat the ordinal data as continuous (e.g., Fokkema, Smits, Kelderman, & Cuijpers, 2013). The log likelihood function of FIML for case *i* is as

$$l_i(\boldsymbol{\theta}) = K_i - \frac{1}{2}\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})_i| - \frac{1}{2}(x_i - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}(\boldsymbol{\theta})_i^{-1}(x_i - \boldsymbol{\mu}_i). \quad (1)$$

K_i is a constant. x_i represents the observed data for case *i*. $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\mu}_i$ represent the model implied covariance matrix and mean vector for case *i*, respectively. The log likelihood for the whole sample is simply a sum of the individual log likelihoods (Arbuckle, 1996, p. 248; Yuan & Bentler, 2000, pp.167–168). The FIML estimates of model parameters can be then obtained by maximizing $l(\boldsymbol{\theta})$,, which can be represented as

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{N} l_i(\boldsymbol{\theta}). \quad (2)$$

The test statistic of FIML is then calculated as

$$T_{\text{FIML}} = -2(l(\widehat{\boldsymbol{\theta}}) - l(\widehat{\boldsymbol{\beta}})), \quad (3)$$

where $l(\widehat{\boldsymbol{\theta}})$is the maximized log likelihood and $l(\widehat{\boldsymbol{\beta}})$is the corresponding maximized log likelihood under a saturated model (Yuan & Bentler, 2000). Based on T_{FIML},, the $\Delta\chi^2$ between two nested models is simply the difference in T_{FIML} between the two models. Say model A is nested within model B, the $\Delta\chi^2$ between the two models is calculated as

$$\Delta\chi^2 = T_{\text{FIML(A)}} - T_{\text{FIML(B)}} = -2(l(\widehat{\boldsymbol{\theta}}_A) - l(\widehat{\boldsymbol{\theta}}_B)). \quad (4)$$

Researchers can include auxiliary variables into the FIML analysis with Graham's saturated model method by correlating all the residual terms of indicators with the auxiliary variables. Including auxiliary variables can help reduce the bias due to missing data (Graham, 2003).

Continuous robust full information maximum likelihood

One limitation of FIML is that it assumes multivariate normality. With nonnormally distributed data, this assumption is violated, which could lead to biased standard error estimates and test statistics (e.g., Teman, 2012). Corrections on the standard error estimates and test statistics from FIML have, thus, been developed (Yuan & Bentler, 2000). Specifically, an adjusting factor (*c*) is multiplied to the T_{FIML} in Equation (3) to mitigate the influence of continuous nonnormality as

$$T_{\text{rFIML}} = c \times T_{\text{FIML}}. \quad (5)$$

For the standard errors, sandwich-type standard errors are calculated using the second derivative of the likelihood function. Detailed information on these corrections can be found in Yuan and Bentler (2000).

With rFIML, $\Delta\chi^2$ cannot be simply calculated by taking the difference between the test statistics for two nested models, but using the formula

$$\Delta\chi^2 = \frac{T_{\text{rFIML\_A}} \times c_A - T_{\text{rFIML\_B}} \times c_B}{c_d}, \quad (6)$$

where c_A and c_B are the corrections factors for model A and model B, respectively. c_d is the correction factor for $\Delta\chi^2$, which is calculated as (See also www.statmodel.com/chidiff.shtml):

$$c_d = \frac{df_A \times c_A - df_B \times c_B}{df_A - df_B}. \quad (7)$$

Similar to FIML, auxiliary variables can be included in rFIML using Graham's saturated model approach; while rFIML may have some corrections for the nonnormality, it is still flawed by assuming the data to be continuous. Past research has shown that this limitation (i.e., treating ordinal as continuous) could result in biased point

estimates and test statistics in SEM for ordinal data (e.g., Li, 2014, 2016; Teman, 2012), even though the biases may not often be substantive (Jia, 2016; Rhemtulla, Brosseau-Liard, & Savalei, 2012).

Full information maximum likelihood methods with probit links and logit links

Besides FIML and rFIML, researchers could also use the full maximum likelihood information methods based on logit (lFIML) or probit links (pFIML) to estimate latent variable models with ordinal missing data. lFIML and pFIML are widely used in the framework of item response theory (Wirth & Edwards, 2007). However, they are not much used in SEM.

Unlike FIML and rFIML which use a linear function to link the latent variables and corresponding indicators, lFIML and pFIML use casewise probit/logit link functions. To illustrate, let y_j be the observed ordinal response for item *j* (*j* = 1….*p*) which has C categories, and η be the score of the latent variable. lFIML assumes that for a participant whose latent variable score is η_i, the probability of this participant endorsing a certain category *k* (*k* = 1, 2, … , C) for item *j* can be written as (see Samejima, 1969)

$$T_j = P(y_j = k|\eta_i) = \begin{cases} 1 - \frac{1}{1+\exp(-Da_j(\eta_i - b_{j1}))} & if \quad k = 1 \\ \frac{1}{1+\exp(-Da_j(\eta_i - b_{j1}))} - \frac{1}{1+\exp(-Da_j(\eta_i - b_{j2}))} & if \quad k = 2 \\ \vdots & \vdots \\ \frac{1}{1+\exp(-Da_j(\eta_i - b_{jc-2}))} - \frac{1}{1+\exp(-Da_j(\eta_i - b_{jc-1}))} & if\, k = C-1 \\ \frac{1}{1+\exp(-Da_j(\eta_i - b_{jc-1}))} & if \quad k = C \end{cases}. \quad (8)$$

D is a scaling constant (typical 1.7) which is used to rescale the results from a logit model to the original scale (i.e., scale in probit model, Wirth & Edwards, 2007). a_j and b_j are the slope and difficulties parameters in a IRT framework, respectively. They can be analytically transformed to parameters that are widely used in SEM framework such as factor loadings (Takane & De Leeuw, 1987; Wirth & Edwards, 2007).

One may replace the logit link in Equation (8) with a probit link to form the equation for a probit model (e.g., Asparouhov & Muthén, 2016). The probability function can be extended to all questions (variables) and all participants in the sample to form the fit function for lFIML(or pFIML). This fit function is then maximized to obtain the estimates.

Superior to the other methods, lFIML or pFIML is capable of handling ordinal missing data using full information maximum likelihood without pretending the data to be continuous. However, they are not limit free. One limitation is that they may not accommodate auxiliary variables in the estimation process with Graham's saturated model approach. The reason is that the estimation process of pFIML and lFIML usually involes numeric intergrations or other computational demanding methods (Wirth & Edwards, 2007). Adding correlations between auxiliary variables and residuals will dramatically increase the dimensionality of numerical integration, which often creates computational problems. Thus, when missingness is determined by auxiliary variables, using lFIML or pFIML without the auxiliary variables will result in a situation that is analogous to missing not at random (MNAR). This situation is referred to as indirect MNAR in Enders (2010). As a result, the bias due to missing data cannot be completely removed by IFIML or pFIML.

WLSMV with pairwise deletion

WLSMV is an extension of the weighted least square estimation method (WLS) for ordinal data. WLS assumes that for each ordinal indicator, there is a normally distributed latent response variate underlying the indicator.

Let y_j be an observed ordinal response for item *j* which has C categories and y_j^* be the latent response variate underlying item *j*; y_j can be created by categorizing y_j^* based on C – 1 thresholds $(\tau_{j,1}, \tau_{j,2}, \ldots . \tau_{j,c-1})$as

$$y_j = \begin{Bmatrix} 1 & if & y_j^* \leq \tau_{j,1} \\ 2 & if & \tau_{j,1} \leq y_j^* \leq \tau_{j,2} \\ \vdots & \vdots & \vdots \\ C-1 & if & \tau_{j,c-2} < y_j^* \leq \tau_{j,c-1} \\ C & if & \tau_{j,c-1} < y_j^* \end{Bmatrix}. \quad (9)$$

A typical estimation process of WLS involves three steps (Muthén, 1984; Muthén, De Toit, & Spisic, 1997; Wirth & Edwards, 2007). First, univariate information of each variable in the sample is used to obtain the maximum likelihood estimates of the sample implied thresholds. Second, polychoric correlations between each pair of the observed indicators are calculated by treating the thresholds obtained in step 1 as fixed (more detail information on the two steps of estimation can be found in Bollen, 1989, pp. 439–443; Olsson, 1979). Third, the estimated thresholds and polychoric correlations are used to form the discrepancy function which is minimized to obtain the estimates for the model parameters. The discrepancy function F_{WLS} can be represented as

$$F_{\text{WLS}} = (\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\theta}))' \mathbf{W}^{-1} (\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\theta})). \quad (10)$$

s is a vector of unique elements in the sample polychoric correlation matrix and thresholds. $\boldsymbol{\sigma}(\boldsymbol{\theta})$ is a vector of model implied polychoric correlations and thresholds. **W** is the weight matrix, which is usually a consistent estimate of the true population asymptotic covariance matrix of **s** (see equation (4) in Muthén et al., 1997). Note that the length or the dimensions of s, $\boldsymbol{\sigma}(\boldsymbol{\theta})$, and **W** are depending on the model complexity (e.g., number of indicators) but not sample size.

The test statistic of WLS can be calculated using the minimized fit function $F_{\text{WLS}}(\widehat{\boldsymbol{\theta}})$as

$$T_{\text{WLS}} = (N-1) \times F_{\text{WLS}}(\widehat{\boldsymbol{\theta}}), df = p* - q, \quad (11)$$

where *N* is sample size, *p** is the number of unique elements in **s**, *q* is the number of estimated parameters.

WLS estimates are consistent and asymptotically follow a normal distribution (Muthén, 1984; Muthén & Satorra, 1995). However, past research showed that WLS required a large sample size, thus, is not practical for typical research in social and behavioral sciences (e.g., Flora & Curran, 2004). To mitigate the problem, a solution is to invert only the diagonal elements of the weight matrix rather than the whole **W** matrix (Muthén et al., 1997). This approach is named diagonal weighted least squares estimation (DWLS). For DWLS, the fit function can be written as

$$F_{\text{DWLS}} = (\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\theta}))' \mathbf{W}_D^{-1} (\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\theta})). \quad (12)$$

$\mathbf{W}_D$ is the diagonalization of the **W** in Equation (10) (Wirth & Edwards, 2007).

Using only diagonal elements of the weight matrix can result in information loss, which can potentially distort the test statistic and standard error estimates (Savalei, 2014). Several methods have been proposed to correct the test statistic and standard error estimates for the information loss. One method corrects the test statistic such that the mean and variance of the test statistic will approximate those of a χ^2 distribution with corresponding degrees of freedom (DiStefano & Morgan, 2014; Muthén et al., 1997). DWLS with this correction

is named WLSMV, which has become most popular. Past research found that WLSMV outperformed the others in terms of the accuracy of χ^2 statistics (DiStefano & Morgan, 2014); it is, thus, considered in our article. With WLSMV, the $\Delta\chi^2$ test statistic between two nested models (say model A nested under model B) is calculated as

$$\Delta\chi^2_{\text{WLSMV}} = N \times m \times (F_{\text{DWLS}}(\widehat{\boldsymbol{\theta}}_A) - F_{\text{DWLS}}(\widehat{\boldsymbol{\theta}}_B)) + l. \quad (13)$$

m is a shift parameter and *l* is a scale parameter. They are used so that the mean and variance of $\Delta\chi^2_{\text{WLSMV}}$ will approximate those of a χ^2distribution with the *df* be equal to the difference in numbers of parameters between models A and B (Asparouhov & Muthén, 2006, 2010b). Note that $F_{\text{DWLS}}(\widehat{\boldsymbol{\theta}}_A)$and $F_{\text{DWLS}}(\widehat{\boldsymbol{\theta}}_B)$in Equation (13) are the minimized values of the fit functions from model A and B but not the χ^2 statistics from the two models. More details on how $\Delta\chi^2_{\text{WLSMV}}$ is calculated can be found in Kite, Johnson, and Chong (2017).

An advantage of WLSMV is that it can handle the multidimensional models that lFIML and pFIML have difficulty to estimate. Auxiliary variables can be also included using Graham's saturated model with WLSMV. However, WLSMV has its own limitation. As we mentioned earlier, in the first two stages of the estimation process of WLSMV, only the univariate and bivariate information in the data are used to calculate sample thresholds and polychoric correlations. WLSMV is not a full information estimation method. As a result, it cannot directly deal with missing data by itself. By default, SEM software (e.g., Mplus) uses traditional deletion methods such as pairwise deletion combined with WLSMV (i.e., WLSMV_PD) to handle the missing data (Asparouhov & Muthén, 2010b). Given that previous studies have found that pairwise deletion could result in an inflated type I error rate for the test statistic in SEM (e.g., Savalei & Bentler, 2005), we expect that this limitation also applies to the $\Delta\chi^2$ from WLSMV_PD.

Purpose of the current research

As can be seen from the description above, none of the methods seems optimal for testing ME/I when ordinal missing data present. Thus, it is not clear which method will perform best and what method(s) would be acceptable in ME/I testing. To our knowledge, no study so far has thoroughly compared the performances of these methods for ME/I testing with ordinal missing data. Thus, empirical researchers have to select a method based on their own' personal preferences without solid justifications (e.g., Fokkema et al., 2013). To fill in this gap in the literature, we conducted a simulation study to compare the relative performances of these methods.

Design of the simulation study

We used a population model similar to that used in Sass et al. (2014) as the baseline model to generate data. This model was a two-group (say groups A and B), single-factor CFA model with ten indicators (see Figure 1). The indicators were all 5-point Likert-type variables. The model parameters included the factor loadings and thresholds for each group. The configural invariance model was identified by fixing the latent factor variances in two groups to be one.

Figure 1. The population model. *Note:* AUX_A and AUX_B represent the auxiliary variables for groups A and B, respectively. LV_A and LV_B represent the latent variables for groups A and B, respectively. $V_{A1} - V_{B10}$ are ordinal indicators and λs are loadings. $\lambda_{A1} - \lambda_{B7}$ are fixed at 0.6. Their thresholds are fixed at τ_n = –1.3, –0.47, 0.47, 1.3 in symmetric conditions and τ_n = –0.253, 0.385, 0.842, 1.282 in asymmetric conditions. In loading noninvariant conditions, $\lambda_{B8} - \lambda_{B10}$ are equal to 0.6 minus a specific value, depending on the amount of noninvariance. Similarly, the thresholds for items 8–10 in group B are equal to their default values (symmetric or asymmetric) minus a specific value in threshold noninvariant conditions, depending on the amount of noninvariance. The residual term of each indicator follows a normal distribution with mean = 0 and variance = 1 – square of the corresponding loading.

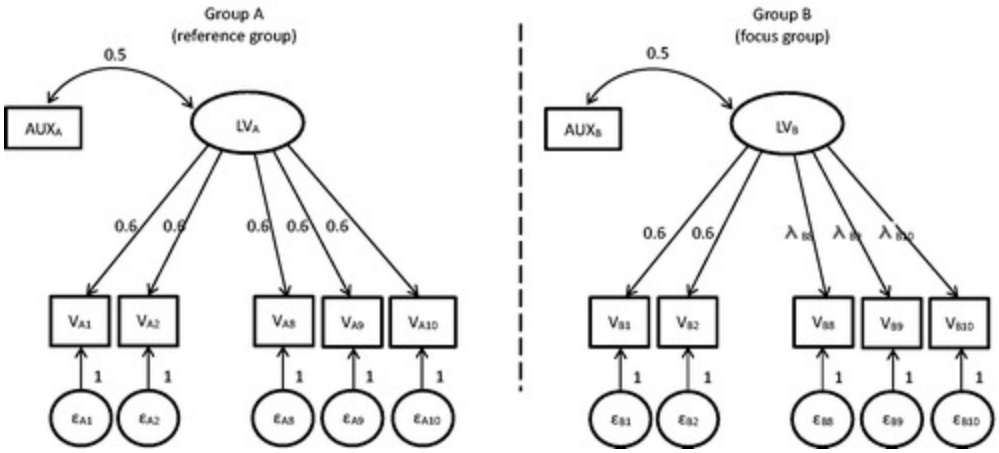

Figure 1.The population model. Note:AUX A and AUX B represent the auxiliary variables for groups A and B, respectively. LV A and LV B represent the latent variables for groups A and B, respectively. V A1 –V B10 are ordinal indicators andks are loadings.k A1 –k B7 are fixed at 0.6. Their thresholds are fixed ats n ¼–1.3,–0.47, 0.47, 1.3 in symmetric conditions ands n ¼–0.253, 0.385, 0.842, 1.282 in asymmetric conditions. In loading noninvariant conditions,k B8 k B10 are equal to 0.6 minus a specific value, depending on the amount of noninvariance. Similarly, the thresholds for items 8–10 in group B are equal to their default values (symmetric or asymmetric) minus a specific value in threshold noninvariant conditions, depending on the amount of noninvariance. The residual term of each indicator follows a normal distribution with mean¼0 and variance¼1–square of the corresponding loading.

The factor loadings in the population model were all set to 0.6, except for the conditions where noninvariance was presented in factor loadings. We varied the thresholds in the current study to generate symmetric and asymmetric thresholds. We also created auxiliary variables (one for each group) that correlated with the latent factor with $r = 0.5$. The auxiliary variable B was used to generate missing data in group B as explained below.

Design factors

The design factors in the study include (1) sample size, (2) location of noninvariance, (3) magnitude of noninvariance, (4) distribution of thresholds, and (5) missing data proportion. The factors were all between replication factors, except for the missing data proportions.

Sample size

The total sample size was varied at three levels: 300 (150 per group), 500 (250 per group), or 1000 (500 per group), representing small, medium, or large sample sizes. These settings are identical to those used in Sass et al. (2014) and similar to those used in other previous studies (e.g., Chen, 2007; Cheung & Rensvold, 2002).

Distribution of thresholds

Population thresholds of items were set to be either symmetric ($\tau_n = –1.3, –0.47, 0.47, 1.3$) or asymmetric ($\tau_n = –0.253, 0.385, 0.842, 1.282$). These settings are identical to those used in Sass et al. (2014).

Location of noninvariance

We set the occurrence of noninvariance to the last three items in group B (i.e., items 8–10 in group B), on either their loadings or thresholds. For convenience, we referred to the conditions containing noninvariant items as noninvariant conditions. Depending on the parameters for which the items are noninvariant, there were loading noninvariant conditions and threshold noninvariant conditions. In contrast, we refer to the conditions where all items were invariant on loadings and thresholds as invariant conditions.

Magnitude of noninvariance

For noninvariant conditions, we varied the magnitude of noninvariance at four levels. Specifically, for a loading noninvariant condition, the loadings of items 8–10 in group B were 0.6 minus 0.2, 0.3, 0.4, or 0.5, representing an increasing magnitude of nonvariance. In a threshold noninvariant condition, 0.2, 0.3, 0.4, or 0.5 was also subtracted from all of the thresholds of items 8–10 in group B. Note that we examined a wider range for the magnitude of noninvariance as compared to previous studies (e.g., Meade & Lautenschlager, 2004; Sass et al., 2014). More details of these parameter settings are presented in Table 1.

Table 1. Model parameters for different amount of noninvariance across conditions.

		Group B				
Parameter	Group A	Amount of noninvariance = 0	0.2	0.3	0.4	0.5
Loadings (items 1–7)	.6	.6	.6	.6	.6	.6
Loadings (items 8–10)	.6	.6	.4	.3	.2	.1
Symmetric						
Thresholds (items 1–7)	(–1.3, –0.47, 0.47, 1.3)	(–1.3, –0.47, 0.47, 1.3)	(–1.3, –0.47, 0.47, 1.3)	(–1.3, –0.47, 0.47, 1.3)	(–1.3, –0.47, 0.47, 1.3)	(–1.3, –0.47, 0.47, 1.3)
Thresholds (items 8–10)	(–1.3, –0.47, 0.47, 1.3)	(–1.3, –0.47, 0.47, 1.3)	(–1.5, –0.67, 0.27, 1.1)	(–1.6, –0.77, 0.17, 1.0)	(–1.7, –0.87, 0.07, 0.9)	(–1.8, –0.97, –0.03, 0.8)
Asymmetric						
Thresholds (items 1–7)	(–0.253, 0.385, 0.842, 1.282)	(–0.253, 0.385, 0.842, 1.282)	(–0.253, 0.385, 0.842, 1.282)	(–0.253, 0.385, 0.842, 1.282)	(–0.253, 0.385, 0.842, 1.282)	(–0.253, 0.385, 0.842, 1.282)
Thresholds (item 8–10)	(–0.253, 0.385, 0.842, 1.282)	(–0.253, 0.385, 0.842, 1.282)	(–0.453, 0.185, 0.642, 1.182)	(–0.553, 0.085, 0.542, 0.982)	(–0.653, – 0.015, 0.442, 0.882)	(–0.753, – 0.115, 0.342, 0.782)

Note: noninvariance occurs only on either loadings or thresholds for items 8–10 in group B.

Missing data proportions

Similar to noninvariance, we imposed missing data on only the last three items in group B. We varied missing data proportions of the three items at three levels: 0%, 30%, and 50%, representing none, moderate, and large proportions of missing data (see also Wu, Jia, & Enders, 2015). Note that even though 50% missing data rate may be high in practice, it could occur in situations such as planned missing data designs, longitudinal studies, or clinical studies.

The missing data were generated as follows. First, the scores of the auxiliary variable B were rank ordered from the smallest to the largest. The probability of missing an item value for individual *i* was then calculated based on the rank order of the auxiliary variable for individual *i* ($rank_i$). For example, let n_b be the number of the observations in group B. The probability of having missing data on the eighth item for an individual *i* in group B is

computed as $1 - \frac{\text{rank}_i}{n_b}$.This probability is then compared to a random number *k* drawing from a uniform distribution, *k* ~ UNIF(0,1). If *k* is less than the calculated probability $k < 1 - \frac{\text{rank}_i}{n_b}$, then individual *i* has a missing observation on the eighth item. This process is continued until the desired percentage (30% or 50%) of missing data is achieved for each of the three items.

In total, there were 144 noninvariant conditions: sample sizes (3) ×× locations of noninvariance (2) ×× magnitudes of the noninvariance (4) ×× distributions of the thresholds (2) ×× missing data proportions (3); and 18 invariant conditions: sample sizes (3) ×× distributions of the thresholds (2) ×× missing data proportions (3). The model parameters of the between replication conditions are presented in Table 1. For each of the conditions, 500 data sets were generated using R. 3.3.1 (R core team, 2016).

Implementations of the methods

The five methods were applied to each of the data sets using Mplus 8.0 (Muthén and Muthén, 1998–2017). ME/I testing was conducted by comparing the configural invariance model to the scalar invariance model using $\Delta\chi^2$ tests. The *df* of the $\Delta\chi^2$ tests obtained from the FIML/rFIML, WLSMV_PD, and lFIML/pFIML were 19, 38, and 48, respectively. Note that these numbers might slightly change (e.g., 38 -> 37) for replications where some categories within items are collapsed due to data sparseness.

When missing data present, the auxiliary variables are included in the analysis using the saturated correlation model for FIML, rFIML, and WLSMV_PD but not for pFIML and lFIML, due to the computational limitations mentioned earlier.

Outcomes

Given that the main focus of the study is on the $\Delta\chi^2$ tests, our primary outcomes were the type I error rates and power associated with the $\Delta\chi^2$ tests obtained from the examined methods. The type I error rate is calculated for each of the invariance conditions, and power is calculated for each of the noninvariance conditions. Both are calculated as the proportion of replications that yield significant $\Delta\chi^2$ tests ($p < .05$). Following Sass et al. (2014), we used 0.030–0.069 as the acceptable range for type I error rates, which was the 95% CI for the nominal level of the type I error rate given the number of replications used in the study.

Our secondary outcomes were the bias of standardized loading estimates and their corresponding standard errors obtained from configural invariance models. We calculated relative biases (RBs) for loading estimates and their standard errors. The RB of a loading estimate was defined as $\frac{(\bar{\theta}_{\text{est}} - \theta_0)}{\theta_0}$,, where θ_0 is the population value of the loading and $\bar{\theta}_{\text{est}}$ is the average of loading estimates across all replications in a given cell of the conditions matrix. The RB of the standard error estimate for a loading was defined as $\frac{\overline{SE} - SE_{\text{emp}}}{SE_{\text{emp}}}$,, where $\overline{SE}$is the average estimated standard error in a given cell. SE_{emp} is the standard deviation of the associated parameter estimates across replications, which is considered as a proxy of the true standard error. Following Flora and Curran (2004) and Rhemtulla et al. (2012), we used |RB| > 0.1 as the threshold for substantially biased estimates.

Results

Nonconvergence and improper solutions

All replications with four FIML methods converged. Only nine out of 81,000 replications had convergence problems with WLSMV_PD. As for the rates of improper solution (i.e., |standardized loadings| > 1) for the loading estimates, FIML and rFIML were more likely than WLSMV_PD, pFIML, and lFIML to produce improper estimates, especially with a small sample size and high missing data rate. The results are summarized in the

Supporting Information. Improper solution rates were also low for all methods (less than 10%), and they decreased as sample size increased. With *N* > 300, they were less than 2%. The replications with improper solutions were excluded from the rest of the analyses.

Type I error rates of $\Delta\chi^2$ tests

The results for type I error rates are summarized in Table 2. In general, all FIML methods outperformed WLSMV_PD in controlling type I error rates at the nominal level when missing data present. The type I error rates from all FIML methods fell into the acceptable range (i.e., .030–.069) in most conditions. FIML and rFIML tended to slightly overcontrol the type I error rates when missing data were present, the sample size was small, and thresholds were asymmetric. In contrast, pFIML and lFIML tended to slightly undercontrol the type I error rates when missing data rates were high.

Table 2. Type I error rate of $\Delta\chi^2$tests.

Method	Thresholds	*N* = 300			*N* = 600			*N* = 1000		
		Complete	30% miss	50% miss	Complete	30% miss	50% miss	Complete	30% miss	50% miss
FIML	Asymmetric	0.043	0.032	0.021	0.054	0.042	0.049	0.046	0.042	0.060
rFIML		0.034	0.032	0.018	0.058	0.042	0.055	0.058	0.048	0.058
WLSMV_PD		0.062	0.072	0.130	0.058	0.094	0.250	0.046	0.108	0.388
pFIML		0.056	0.060	0.066	0.060	0.064	0.070	0.052	0.064	0.074
lFIML		0.050	0.054	0.064	0.058	0.058	0.076	0.056	0.072	0.068
FIML	Symmetric	0.04	0.036	0.046	0.028	0.026	0.032	0.032	0.032	0.054
rFIML		0.056	0.050	0.054	0.040	0.040	0.040	0.056	0.044	0.064
WLSMV_PD		0.062	0.088	0.180	0.036	0.092	0.292	0.052	0.122	0.550
pFIML		0.06	0.058	0.074	0.066	0.052	0.064	0.054	0.078	0.088
lFIML		0.06	0.054	0.070	0.060	0.054	0.062	0.046	0.070	0.076

Note. The values fell out of the acceptable range (0.030 – 0.069) were highlighted. FIML = continuous FIML. rFIML = continuous robust FIML, pFIML = FIML with a probit link, lFIML = FIML with a logit link.

The type I error rates from WLSMV_PD were highly influenced by the missing data rate. With complete data, the type I error rates from WLSMV_PD were all acceptable (.042–.058). However, when missing data presented, the type I error rates from WLSMV_PD were inflated (> .069), especially when the sample size and missing data rate were both large. For example, with 50% missing data and *N* = 1000, the type I error rate could be as high as 0.55.

Power of $\Delta\chi^2$ tests

The results for power to detect noninvariance in loadings with symmetric thresholds are plotted in Figure 2. The results for thresholds noninvariant conditions and loadings noninvariant conditions with asymmetric thresholds are reported in the Supporting Information, given that the methods did not differ for these conditions. As shown in Figure 2, when the sample size was large (*n* = 1000), all methods had sufficient power (> 0.8) to detect noninvariance, except for very few conditions where the amount of noninvariance was small (amount of noninvariance = 0.2). Similarly, when the amount of noninvariance was sufficiently large (≥ .40), all methods had sufficient power to detect noninvariance regardless of sample size. In addition, holding the other factors constant, an increase in the missing data rate resulted in a decrease in the power of $\Delta\chi^2$ for all methods.

Figure 2. Power of the $\Delta\chi^2$ tests on detecting noninvariant loadings when thresholds are symmetric. *Note:* FIML = continuous full information likelihood method, rFIML = robust continuous full information likelihood method, W_PD = weighted least squares means and variance adjusted estimators plus pairwise deletion, pFIML = FIML with a probit link, lFIML = FIML with a logit link. The power of WLSMV_PD in missing data conditions are spurious given its highly inflated type I error rates shown in these conditions in Table 2.

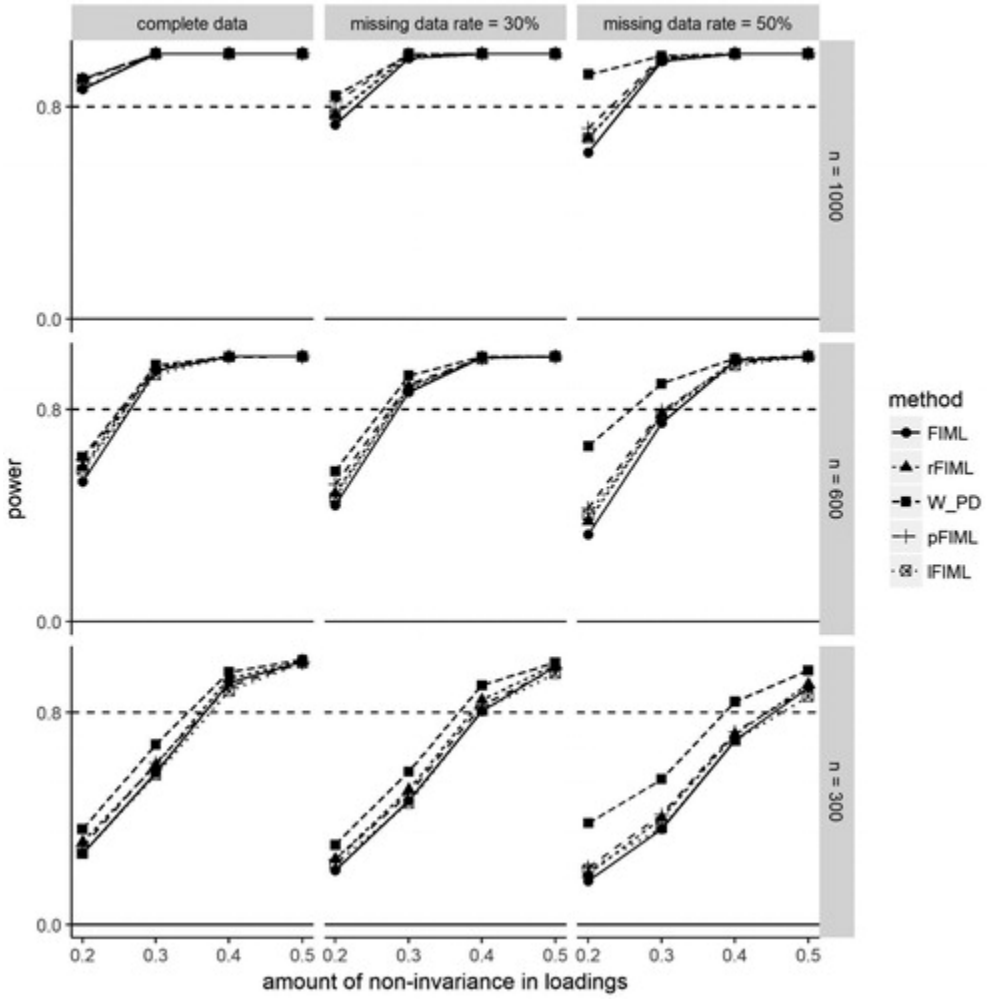

Figure 2.Power of theDv 2 tests on detecting noninvariant loadings when thresholds are symmetric. Note:FIML¼continuous full information likelihood method, rFIML¼robust continuous full information likelihood method, W_PD¼weighted least squares means and variance adjusted estimators plus pairwise deletion, pFIML¼FIML with a probit link, lFIML¼FIML with a logit link. The power of WLSMV_PD in missing data conditions are spurious given its highly inflated type I error rates shown in these conditions inTable 2. MULTIVARIATE BEHAVIORAL RESEARCH9

All FIML methods had comparable power rates across all conditions, and differed from WLSMV_PD. As shown in Figure 2, WLSMV_PD had the highest power* (suspicious power) to detect noninvariance when the missing data rate was high, and either the sample size or amount of noninvariance was not large. However, given the inflated type I error rates associated with WLSMV_PD, these power* rates were not really meaningful.

Relative biases of loading estimates

For ease of presentation, we separated the items into two groups. The first group contained complete items for which the data were always complete. These items included all items in group A and items 1–7 in group B. The second group contained three items that had missing data in some of the conditions (i.e., items 8–10 in group B). We refer to these items as incomplete items. Given that the relative performances of the methods were similar in incomplete and complete items, we only presented the mean relative biases (MRBs) for incomplete items in Figure 3. The MRBs of loading estimates for complete items can be found in the Supporting Information. In addition, because the location of noninvariance and amount of noninvariance did not affect the relative performance of the methods, we collapsed the results across these two factors in Figure 3 as well.

Figure 3. Absolute mean relative biases across incomplete items in group B. *Note:* FIML = continuous full information likelihood method, rFIML = robust continuous full information likelihood method, W_PD = weighted least squares means and variance adjusted estimators plus pairwise deletion, pFIML = FIML with a probit link, lFIML = FIML with a logit link. FIML and rFIML had identical point estimates. Thus, their MRBs are completely overlapped in this figure.

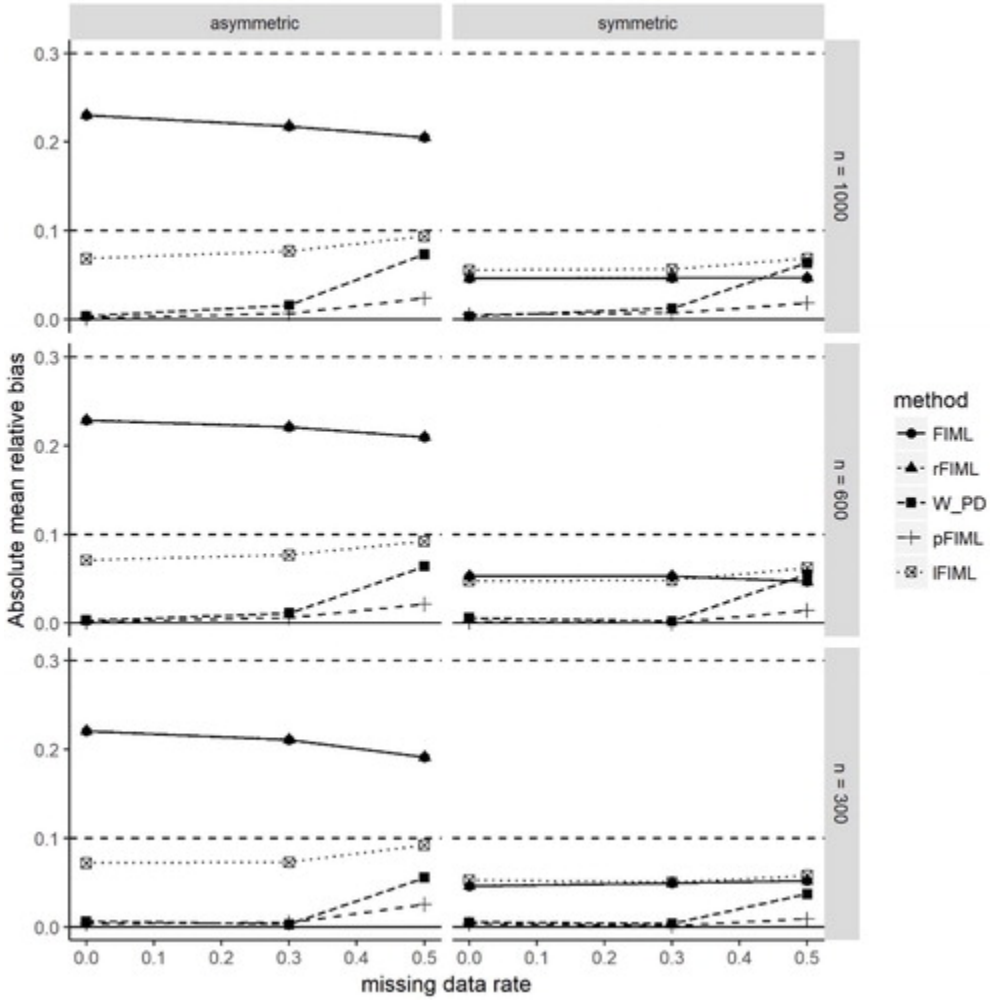

Figure 3.Absolute mean relative biases across incomplete items in group B. Note:FIML¼continuous full information likelihood method, rFIML¼robust continuous full information likelihood method, W_PD¼weighted least squares means and variance adjusted estimators plus pairwise deletion, pFIML¼FIML with a probit link, lFIML¼FIML with a logit link. FIML and rFIML had identical point estimates. Thus, their MRBs are completely overlapped in this figure.

As can be seen in Figure 3, the MRBs of loadings from the FIML and rFIML were mainly affected by the asymmetry of thresholds. When the thresholds were symmetric, the MRBs from the methods were all within the acceptable range (i.e., |RB| < 0.1), regardless of the missing data proportions. However, when the thresholds were asymmetric, the MRBs from FIML and rFIML became substantial. In contrast, the MRBs from the methods that account for ordinal nature of the data (i.e., WLSMV_PD, pFIML, and lFIML) were all acceptable (only slighty affected by missing data rates, |MRB| < 0.1), especially for the loading estimates obtained from pFIML.

Relative biases in standard error estimates

As for the standard errors (SEs) for loading estimates, the MRBs from all methods were mostly within the acceptable range (|RB| < 0.1) for complete items, so were they for incomplete items with symmetric thresholds. Thus, for simplicity, we only presented the results for incomplete items with asymmetric thresholds in Table 3.

The other results can be found in the Supporting Information. As shown in Table 3, SEs obtained from FIML were always substantively biased in thresholds noninvariant conditions. The SEs from rFIML and WLSMV_PD were biased only in conditions where sample sizes were small (*n* = 300) and missing rates were high (50%). In comparison, the SEs from lFIML and pFIML were accurate across all conditions in the current study.

Table 3. Mean relative biases of standard errors for loadings across incomplete items in group B with asymmetric thresholds.

Estimator	DIF_T	DIF_L	*N* = 300			*N* = 600			*N* = 1000		
			Complete	30% miss	50% miss	Complete	30% miss	50% miss	Complete	30% miss	50% miss
FIML	0	0	0.023	0.059	0.076	0.043	0.052	0.090	0.009	0.018	0.020
	0.2	0	0.115	0.185	0.185	0.105	0.121	0.173	0.039	0.040	0.060
	0.3	0	0.169	0.188	0.267	0.106	0.107	0.148	0.102	0.106	0.107
	0.4	0	0.189	0.182	0.229	0.171	0.184	0.162	0.164	0.172	0.187
	0.5	0	0.222	0.239	0.223	0.141	0.156	0.134	0.117	0.122	0.105
	0	0.2	–0.030	–0.009	–0.037	0.036	0.037	0.047	0.014	0.014	0.007
	0	0.3	–0.045	–0.043	–0.021	–0.038	–0.030	–0.042	0.012	0.012	0.002
	0	0.4	–0.017	0.004	–0.035	–0.041	–0.046	–0.026	–0.032	–0.025	–0.020
	0	0.5	–0.041	–0.027	–0.052	–0.024	–0.017	–0.034	0.015	0.003	–0.004
rFIML	0	0	0.025	0.049	0.052	0.030	0.020	0.041	–0.009	–0.018	–0.034
	0.2	0	0.056	0.118	0.118	0.036	0.042	0.085	–0.032	–0.041	–0.030
	0.3	0	0.086	0.105	0.177	0.013	0.009	0.045	0.005	0.002	0.000
	0.4	0	0.084	0.081	0.134	0.054	0.063	0.048	0.043	0.047	0.062
	0.5	0	0.099	0.121	0.122	0.010	0.026	0.018	–0.015	–0.010	–0.018
	0	0.2	0.007	0.022	–0.014	0.060	0.051	0.050	0.032	0.023	0.007
	0	0.3	–0.001	–0.006	0.012	–0.010	–0.007	–0.026	0.035	0.029	0.011
	0	0.4	0.028	0.045	0.001	–0.012	–0.022	–0.005	–0.008	–0.005	–0.005
	0	0.5	–0.006	0.006	–0.023	0.001	0.006	–0.012	0.033	0.018	0.008
W_PD	0	0	–0.052	–0.067	–0.145	–0.002	–0.015	–0.035	–0.030	–0.047	–0.067
	0.2	0	–0.038	–0.038	–0.111	–0.005	–0.016	–0.034	–0.047	–0.066	–0.062
	0.3	0	–0.044	–0.083	–0.087	–0.038	–0.051	–0.057	–0.017	–0.021	–0.034
	0.4	0	–0.048	–0.075	–0.072	–0.001	0.011	–0.047	0.018	0.001	–0.014
	0.5	0	–0.043	–0.066	–0.083	–0.021	–0.037	–0.082	–0.035	–0.037	–0.043
	0	0.2	–0.074	–0.082	–0.142	0.011	–0.019	–0.025	0.009	0.006	–0.026
	0	0.3	–0.077	–0.086	–0.116	–0.052	–0.052	–0.078	0.006	–0.005	–0.018

	0	0.4	–0.059	–0.060	–0.107	–0.060	–0.071	–0.058	–0.030	–0.029	–0.040
	0	0.5	–0.080	–0.085	–0.139	–0.040	–0.046	–0.076	0.010	–0.005	–0.018
pFIML	0	0	0.005	–0.020	–0.028	–0.013	–0.019	–0.014	–0.008	–0.007	–0.042
	0.2	0	–0.005	–0.003	–0.041	–0.026	–0.018	–0.032	0.001	0.022	0.005
	0.3	0	–0.013	–0.021	–0.034	–0.021	0.006	–0.024	–0.001	0.014	0.002
	0.4	0	0.023	–0.013	–0.032	–0.031	–0.045	–0.040	–0.001	–0.005	–0.018
	0.5	0	–0.002	0.001	–0.030	–0.030	–0.021	–0.022	–0.010	–0.007	0.002
	0	0.2	–0.015	–0.028	0.002	0.010	0.002	–0.014	0.012	0.025	–0.015
	0	0.3	0.000	0.015	–0.023	–0.017	–0.016	–0.005	–0.001	–0.009	0.004
	0	0.4	–0.012	–0.023	–0.037	0.008	–0.005	0.021	0.010	–0.011	0.000
	0	0.5	–0.035	–0.042	–0.027	–0.010	–0.014	–0.029	–0.010	–0.006	–0.020
lFIML	0	0	0.020	–0.005	–0.020	0.004	–0.007	–0.008	0.008	0.006	–0.031
	0.2	0	0.010	0.006	–0.030	–0.015	–0.012	–0.030	0.014	0.032	0.020
	0.3	0	0.002	–0.010	–0.026	–0.004	0.022	–0.011	0.013	0.028	0.005
	0.4	0	0.032	–0.001	–0.028	–0.009	–0.029	–0.026	0.009	0.000	–0.018
	0.5	0	0.006	0.007	–0.025	–0.020	–0.011	–0.012	0.005	0.005	0.010
	0	0.2	–0.012	–0.022	–0.001	0.014	0.012	–0.007	0.017	0.033	–0.014
	0	0.3	0.003	0.021	–0.017	–0.015	–0.014	0.000	0.004	–0.008	0.006
	0	0.4	–0.014	–0.027	–0.039	0.011	–0.003	0.025	0.015	–0.006	0.006
	0	0.5	–0.030	–0.046	–0.026	–0.012	–0.015	–0.028	–0.004	0.000	–0.023

Note: rFIML = robust FIML, W_PD = mean and variance adjusted weight least squared with pairwise deletion, pFIML = FIML with a probit link, and lFIML = FIML with a logit link. DIF_T = amount of noninvariance in thresholds, DIF_L = amount of noninvariance in loadings. The absolute values above 0.100 were highlighted.

Empirical example

An empirical example is also used to further demonstrate the relative performances between the examined methods. The data for the empirical example were collected through World Health Organization Quality-of-Life Scale (WHOQOL-BREF) by Chen and Yao (2015). For simplicity, we only used the data from the psychological domain of WHOQOL-BREF, which is a single-factor subscale with 6 five-point Likert-type scale items. We tested ME/I of the scale across gender (158 males and 240 females). Given that the original data were almost complete, we imposed missing data on the last three items in the domain for female participants, based on participants' scores on another measure of general quality of life (auxiliary variable). We tried two scenarios: (1)

participants with lower general quality of life are more likely to have missing data, and (2) participants with higher general quality of life are more likely to have missing data. The missing data rates were set at two levels: 30% or 50%.

The results were very similar between the two scenarios. Thus, only the results for the former are presented in Table 4. One can notice that the $\Delta\chi^2$ statistics obtained from WLSMV_PD were quite sensitive to missing data. Its values quickly increased (increased by 50%) as the missing data rates increased; In contrast, missing data only had trivial influence on the $\Delta\chi^2$ statistics obtained from the four FIML methods.

Table 4. $\Delta\chi^2$test statistics between configural and scalar invariance models across gender on the psychological domain subscale of the WHOQOL-BREF.

		$\Delta\chi^2$ statistics (percentage of inflation versus complete data)	
Methods (df)	Complete data	30% missing	50% missing
FIML (10)	11.151	12.960 (16.2%)	11.472 (2.8%)
rFIML (10)	10.481	12.211 (16.5%)	10.215(–2.6%)
WLSMV_PD (22)	21.227	29.550 (39.2%)	32.762 (54.3%)
lFIML (28)	43.595	47.051 (7.9%)	45.460 (4.2%)
pFIML (28)	44.161	47.204 (6.8%)	44.930 (1.7%)

Conclusions and discussion

In this study, we compared five methods that may be used for ME/I testing with ordinal missing data in recovering the $\Delta\chi^2$ test statistic, loading estimates and standard errors for loading estimates. In the following, we summarize and discuss the major findings.

Type I error rate of $\Delta\chi^2$ tests

WLSMV_PD could lead to highly inflated type I error rates with the presence of missing data. In contrast, the four FIML methods had a much better control of type I error rates. These findings are consistent with previous simulation studies focused on continuous missing data and χ^2 (Marsh, 1998; Savalei & Bentler, 2005). There are two explanations for the inflated type I error rates for $\Delta\chi^2$ tests of WLSMV_PD. First, as aforementioned, WLSMV treats the thresholds and correlation matrix calculated based on the PD as if they are from complete data, so uncertainty in $\Delta\chi^2$ due to missing data is not accounted for. Second, PD could result in a nonuniform sample decrease across summary statistics, which distorts the χ^2 test statistic (Bollen, 1989; Kaplan, 2014).

Power of the $\Delta\chi^2$ tests

As aforementioned, the high power rates associated with WLSMV_PD are not really meaningful. The four FIML methods all had sufficient power to detect noninvariance when the sample sizes or the amounts of noninvariance were moderate or large. rFIML, pFIML, and lFIML slightly outperformed FIML by having higher power to detect noninvariance under some conditions where the amounts of noninvariance are not large.

Loading estimates

All methods produced accurate loading estimates when the thresholds were symmetric. However, when the thresholds were asymmetric, the loading estimates obtained from FIML/rFIML were always biased, even in complete data conditions. In contrast, the loading estimates from WLSMV_PD, lFIML, and pFIML were accurate and only slighted affected by missing rates. These results are probably due to the fact that FIML and rFIML do not account for the discrete nature of the ordinal data. Similar results have also been found in previous studies (e.g., Rhemtulla et al., 2012).

Standard errors

The standard errors obtained from all methods were quite accurate for complete items. However, the methods performed differently for incomplete items with asymmetric thresholds; lFIML and pFIML turned out to be the two best methods, given that only they produced accurate standard errors regardless of the missing data rates and sample sizes. In contrast, the standard errors from FIML, rFIML, and WLSMV_PD were substantively biased in conditions where the sample size was small and missing data rates were high. Overall, FIML had the worst performance in terms of standard errors given that it does not have any adjustment built in for nonnormality due to ordinal data. It produced biased standard errors in almost all conditions where thresholds were asymmetric.

Practical recommendations

Based on the results of our study, we recommend lFIML, pFIML, and rFIML for the $\Delta\chi^2$ tests in ME/I testing. These methods were capable of controlling the type I error rate at acceptable levels while still maintaining sufficient power to detect noninvariance in loadings and thresholds. Comparing to rFIML, iFIM, and pFIML were also able to provide accurate standard errors and loading estimates across the conditions, even without including auxiliary variables. Thus, when $\Delta\chi^2$ tests, parameter estimates, and standard errors are jointly considered, pFIML and rFIML are the best choices. The only limitation of pFIML and lFIML is that they may not run for complicated models (e.g., CFA model with many factors and correlated residuals). In this case, rFIML is a good alternative, except that researchers should be cautious about the parameter estimates and standard errors produced by rFIML if the indicator distributions are asymmetric. The WLSMV_PD should be avoided if a substantial proportion of ordinal missing data exist.

Limitations and future directions

Like any other simulation studies, we could not examine all possible conditions of interest to researchers. For example, similar to Sass et al. (2014), we limited our simulation to five-point ordinal data. Previous studies have shown that reducing the number of categories per item could affect the χ^2 test statistic when treating ordinal data as continuous with continuous maximum likelihood estimator (ML) and robust ML (e.g., Rhemtulla et al., 2012). It will be interesting to see whether this conclusion applies to $\Delta\chi^2$ tests.

Another limitation is that we assume that researchers have the correctly specified configural invariance model. Testing configural invariance could be challenging because its χ^2 statistic can be rejected for one of the following two reasons, or both: (1) the factor structure is not identical across groups, and (2) model misspecification is not relevant to measurement invariance (Jorgensen, Kite, Chen, & Short, 2018). It would be worthwhile to examine whether model misspecification could affect the relative performance of the strategies.

Furthermore, we assume that latent factors are normally distributed, which may not necessarily be true in practice. Suh (2015) found that nonnormally distributed latent variables can affect the performance of $\Delta\chi^2$ tests obtained from WLSMV and maximum likelihood methods with probit/logit link in the context of ME/I testing. In addition, nonnormality can confound with the missing data mechanism that researchers use to generate missing data to affect the performance of the methods. For example, in the current study, we simulated missing data such that missing data were more likely to occur with higher auxiliary scores. Graham (2003) considered this way of generating missing data as a linear missing at random (MAR). It is also possible to generate missing data on both sides of the variables and create nonlinear (or convex) MAR (e.g., imposing missingness to participants with high and low auxiliary scores). When data are normally distributed, Graham found that FIML worked fine for either linear or nonlinear MAR. However, Savalei and Falk (2014) found that when data were nonnormally distributed, the performance of rFIML was sensitive to the different MAR mechanisms. Future research can further investigate the joint effects of nonnormally distributed data and different missing data mechanisms on ME/I testing.

Finally, we found that the performance of lFIML and pFIML was not affected by their limitation of not being able to include the auxiliary variable. This finding may not be always true if there is a stronger relationship between the auxiliary variable and missingness on the model variables, and/or if there are more incomplete variables in the model. Future research is warranted to identify the conditions, if they exist, where the performance of lFIML and pFIML may be compromised by this limitation.

Article information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not supported.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The authors would like to thank two anonymous reviewers, editors, and Dr. Po-Hsien Huang for their comments on prior versions of this manuscript. The first version of this article was submitted when the first author was at the University of Kansas; while the revisions were done after the first author has moved to the University of Texas, Rio Grande Valley. We would like to thank both universities for their support. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions is not intended and should not be inferred.

References

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. *Advanced Structural Equation Modeling: Issues and Techniques*, *243*, 277.

Asparouhov, T., & Muthén, B. O. (2006). Robust chi square difference testing with mean and variance adjusted test statistics. Mplus Web Notes No. 10. Retrieved from http://www.statmodel.com/download/webnotes/webnote10.pdf.

Asparouhov, T., & Muthén, B. O. (2010a). Multiple imputation with Mplus. Mplus Technical Appendix. Retrieved from http://www.statmodel.com/download/Imputations7.pdf.

Asparouhov, T., & Muthén, B. O. (2010b). Simple second order chi-square correction. Mplus Technical Appendix. Retrieved from https://www.statmodel.com/download/WLSMV_new_chi21.pdf

Asparouhov, T., & Muthén, B. O. (2016). IRT in Mplus. Mplus Technical report. Retrieved from http://www.statmodel.com/download/MplusIRT.pdf

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464–504. doi: 10.1080/10705510701301834

Chen, P. Y., & Yao, G. (2015). Measuring quality of life with fuzzy numbers: In the perspectives of reliability, validity, measurement invariance, and feasibility. *Quality of Life Research*, *24*(4), 781–785. doi: 10.1007/s11136-014-0816-3

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255. doi: 10.1207/S15328007SEM0902_5

DiStefano, C., & Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling*, *21*(3), 425–438. doi: 10.1080/10705511.2014.915373

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466. doi: 10.1037/1082-989X.9.4.466

Fokkema, M., Smits, N., Kelderman, H., & Cuijpers, P. (2013). Response shifts in mental health interventions: An illustration of longitudinal measurement invariance. *Psychological Assessment*, *25*(2), 520. doi: 10.1037/a0031669

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, *10*(1), 80–100. doi: 10.1207/S15328007SEM1001_4

Jia, F. (2016). *Methods for handling missing non-normal data in structural equation modeling* (Unpublished doctoral dissertation). University of Kansas, Lawrence, KS.

Jorgensen, T. D., Kite, B. A., Chen, P. Y., & Short, S. D. (2018). Permutation randomization methods for testing measurement equivalence and detecting differential item functioning in multiple-group confirmatory factor analysis. *Psychological Methods*, *23*(4), 708–728. doi: 10.1037/met0000152

Kaplan, D. (2014). *Bayesian statistics for the social sciences* New York, NY: Guilford.

Kite, B. A., Johnson, P. E., & Chong, X. (2017). Replicating the Mplus DIFFTEST procedure: An R function to reproduce nested model comparisons. Retrieve from http://crmda.dept.ku.edu/guides/44.difftest/44.difftest.html.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. New York, NY: Guilford.

Li, C.-H. (2014). *The performance of MLR, USLMV, and WLSMV estimation in structural regression models with ordinal variables* (Unpublished doctoral dissertation). Michigan State University, East Lansing, MI.

Li, C. H. (2016). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods*, *21*(3), 369–387. doi: 10.1037/met0000093

Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, *22*(3), 486–506. doi: 10.1037/met0000075

Marsh, H. W. (1998). Pairwise deletion for missing data in structural equation models: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. *Structural Equation Modeling*, *5*(1), 22–36. doi: 10.1080/10705519809540087

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*(4), 361–388. doi: 10.1177/1094428104268027

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. doi: 10.1007/BF02294825

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. doi: 10.1007/BF02294210

Muthén, B. O., De Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes (Unpublished Technical Report). Retrieved from https://www.statmodel.com/download/Article_075.pdf

Muthén, B. O., & Satorra, A. (1995). Technical aspects of Muthén's LISCOMP approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika*, *60*(4), 489–503. doi: 10.1007/BF02294325

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Author.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443–460. doi: 10.1007/BF02296207

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. doi: 10.1037/a0029315

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No 17.

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling*, *21*(2), 167–180. doi: 10.1080/10705511.2014.882658

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514. doi: 10.1007/BF02296192

Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling*, *21*(1), 149–160. doi: 10.1080/10705511.2013.824793

Savalei, V., & Bentler, P. M. (2005). A statistically justified pairwise ML method for incomplete nonnormal data: A comparison with direct ML and pairwise ADF. *Structural Equation Modeling*, *12*(2), 183–214. doi: 10.1207/s15328007sem1202_1

Savalei, V., & Falk, C. F. (2014). Robust two-stage approach outperforms robust full information maximum likelihood with incomplete nonnormal data. *Structural Equation Modeling*, *21*(2), 280–302. doi: 10.1080/10705511.2014.882692

Suh, Y. (2015). The performance of maximum likelihood and weighted least square mean and variance adjusted estimators in testing differential item functioning with nonnormal trait distributions. *Structural Equation Modeling*, *22*(4), 568–580. doi: 10.1080/10705511.2014.937669

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408. doi: 10.1007/BF02294363

Teman, E. D. (2012). *The performance of multiple imputation and full information maximum likelihood for missing ordinal data in structural equation models* Ann Arbor, MI: ProQuest.

Widaman, K. F., Grimm, K. J., Early, D. R., Robins, R. W., & Conger, R. D. (2013). Investigating factorial invariance of latent variables across populations when manifest variables are missing completely. *Structural Equation Modeling*, *20*(3), 384–408. doi: 10.1080/10705511.2013.797819

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58. doi: 10.1037/1082-989X.12.1.58

Wu, W., Jia, F., & Enders, C. (2015). A comparison of imputation strategies for ordinal missing data on Likert scale variables. *Multivariate Behavioral Research*, *50*(5), 484–503. doi: 10.1080/00273171.2015.1022644

Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*(1), 165–200. doi: 10.1111/0081-1750.00078